# Everything You Always Wanted to Know About AI Safety (But Were Afraid to Ask)

David Krueger

# About me

# My background

- Deep Learning since 2013
- AI Safety since before that
- Mila (grad) → Cambridge (prof) → Mila (prof)

**I’ve been advocating for AI Safety within the machine learning community for >10 years.**

# My research

- Deep learning
  - **Key Question: How does generalization work in deep learning?**
    - “a closer look at memorization in deep networks” - 2017
    - “Out-of-Distribution Generalization via Risk Extrapolation (REx)” - 2021
- AI Safety
  - **Key Question: (how and why) are AI “agents” dangerous?**
    - “Hidden Incentives for Auto-Induced Distributional Shift” - 2020
    - “goal misgeneralization in deep reinforcement learning” - 2021
    - “defining and characterizing reward hacking” - 2022
    - “harms from increasingly agentic algorithmic systems” - 2023
    - “implicit meta-learning may lead language models to trust more reliable sources” - 2024

# AI Safety: The Basic Case

# A brief case for AI x-risk

1. Instrumental goals

***"The ends justify the means"***

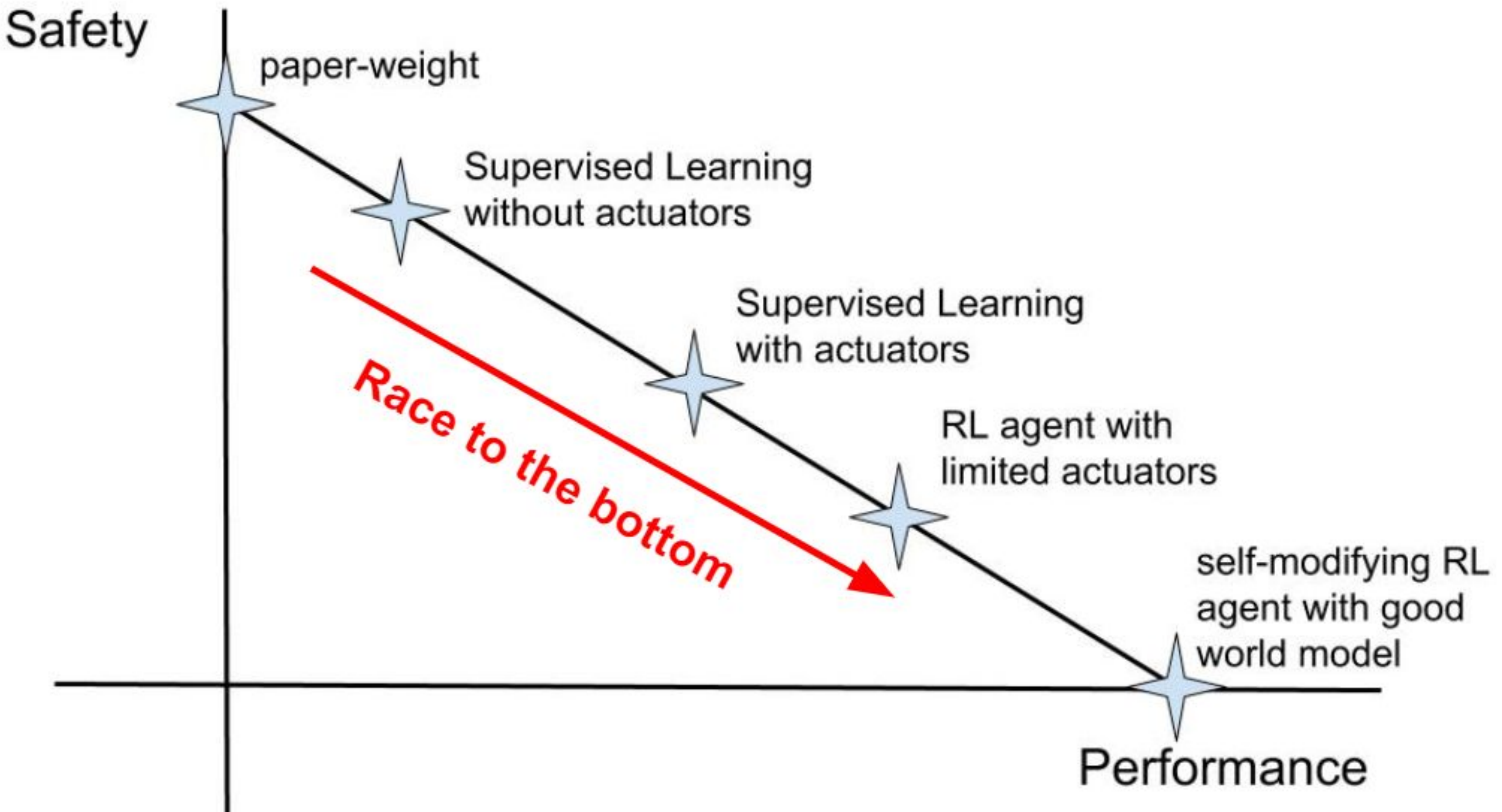
2. Goodhart's law

***"You get what you measure"***

3. Safety-performance trade-offs: **For greater safety…**
    - a. **Do lots of testing**
    - b. **Keep a human-in-the-loop**
    - c. **Require interpretability**
    - d. **No long-term planning (remove instrumental goals)**
    - e. **Limit sensors and actuators**
    - f. **Keep the system narrow**
    - g. **Hard-code safety constraints**

"Gorilla problem"

# Safety - Performance Trade-off



# Companies are trying to build artificial people

- AI companies are trying to take everyone’s jobs
- To do that, they need to make robots
- And the robots need to be “agents”

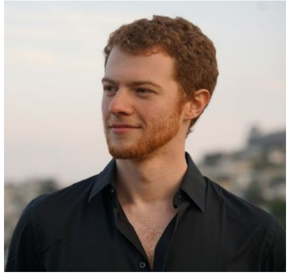
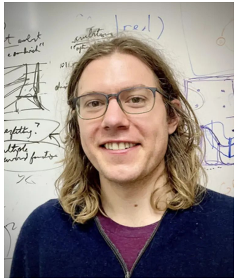
So by default, we’re going to end up with a bunch of autonomous robots running around pursuing long-term goals.

**“A Cambrian Explosion of Artificial Life”™**

# Gradual Disempowerment



Jan Kulveit, Raymond Douglas, Nora Ammann


Deger Turan, David Krueger, David Duvenaud

https://gradual-disempowerment.ai/

# Traditional View of AI Risk:

- Sudden jump in capabilities
- Misuse (cyberattacks, bioweapons)
- Autonomous systems betray us

# Gradual Disempowerment

- Human cognition replaced everywhere
- Could happen incrementally
- Driven by local incentives
- Hard to coordinate against

**David Krueger**
@DavidSKrueger

AI automation is NOT just an economic issue!

Labor doesn't just give you money, it also gives you power.  When the world doesn't rely on people power anymore, the risk of oppression goes up.

Right now, popular uprisings can and do regularly overthrow oppressive governments.

A big part of that is because the military and police are made up of people -- people who can change sides or stand down when the alternative is too risky or abhorrent to them.

When the use of force at scale no longer requires human labor, we could be in big trouble.

12:31 PM · Sep 6, 2025 · **17.3K** Views

# A rough vignette

- Work becomes more intense, then gradually becomes monitoring / make-work.
- Your friends and family get really into their AI companions
- You get really into your AI companions
- It’s harder to tell what’s going on
- …
- Eventually you’re forced to be uploaded
    - unclear if you’ll be run much

# Three 'Ingredients'

## 1. Social systems are 'contingently aligned'

Economy, states, culture, etc. have tended to produce pro-human outcomes.

This is not inherent to these systems, but due to the presence of:

- ***Explicit* alignment mechanisms**, e.g. consumer choice, voting
- ***Implicit* alignment mechanisms**, e.g. dependency on human participation/labour/cognition

# Three 'Ingredients'

## 2. Selective pressures at every scale

AI systems will increasingly replace humans in critical social functions, either by **deliberate choice**, or **through selection pressure.**

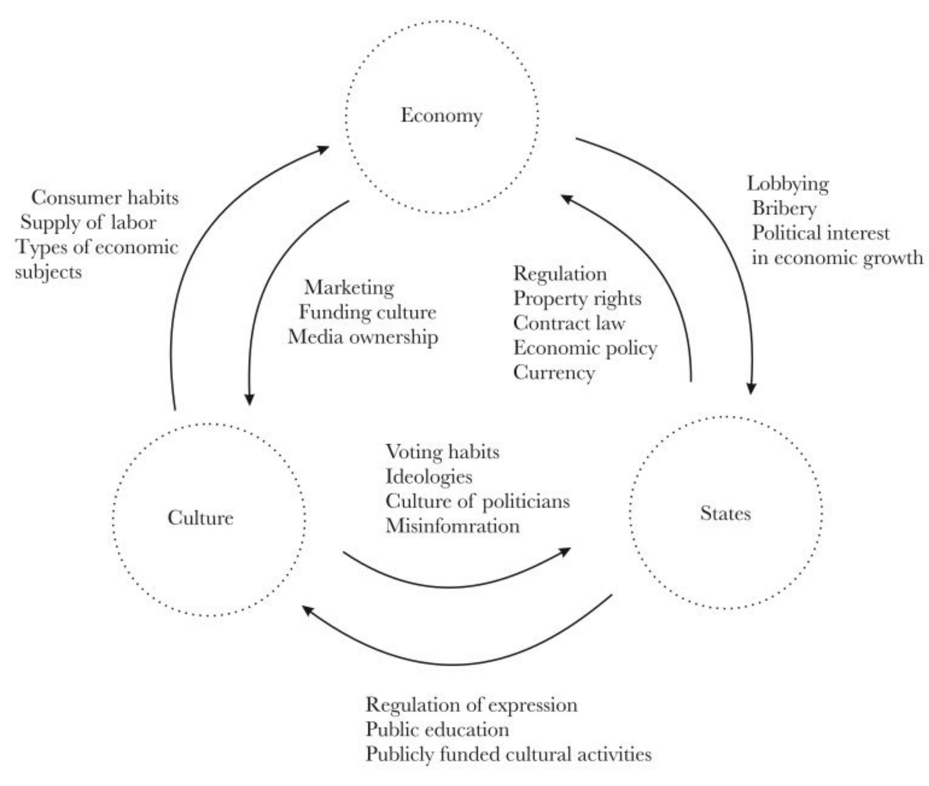
This will undermine or break these (explicit & implicit) alignment mechanisms.

# Three 'Ingredients'

## 3. 'Wicked' System Interactions

(Naive) interventions to maintain human empowerment over one area tend to undermine empowerment in another area.

- **UBI?** Reduces state's accountability vis a vis its (human) citizen. Increases state's leverage over them.
- **Ban or restrict AI use in economy & culture?** Competition & black markets. State overreach. Eroding institutional legitimacy.
- **Strengthen citizen's power over state?** Makes state more susceptible to shifts in culture, including ones driven & shaped by AI.



# Big Concepts in AI Safety

# (Existential) Safety

- X-risk
  - Not just extinction
- Longtermism
- Effective Altruism

# AI Alignment

Technical

AI Alignment = How to get AI systems to do **what we want**

AI Alignment = How to get AI systems **to try** to do **what we want**

AI Alignment = A rebranding of **“AI (existential) safety”**... A community of people trying to reduce the chance of AI leading to premature human extinction.

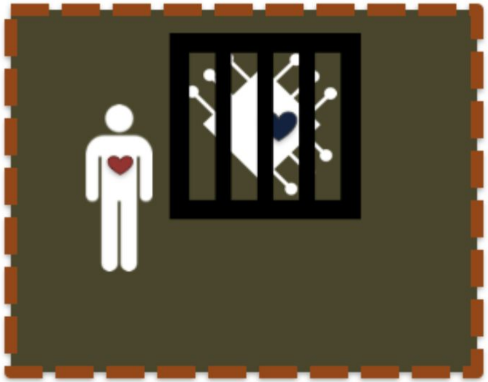
Socio-Technical

# Alignment vs. capabilities

"AI Control"

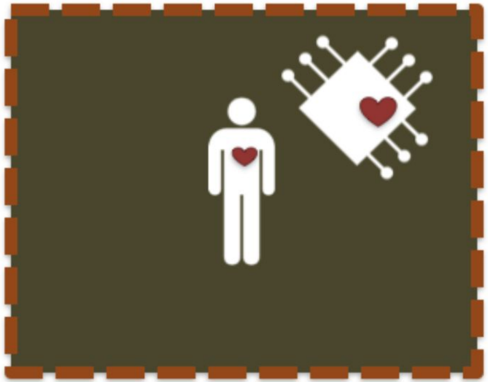
Two Technical Approaches to AI Alignment:

Capability Control



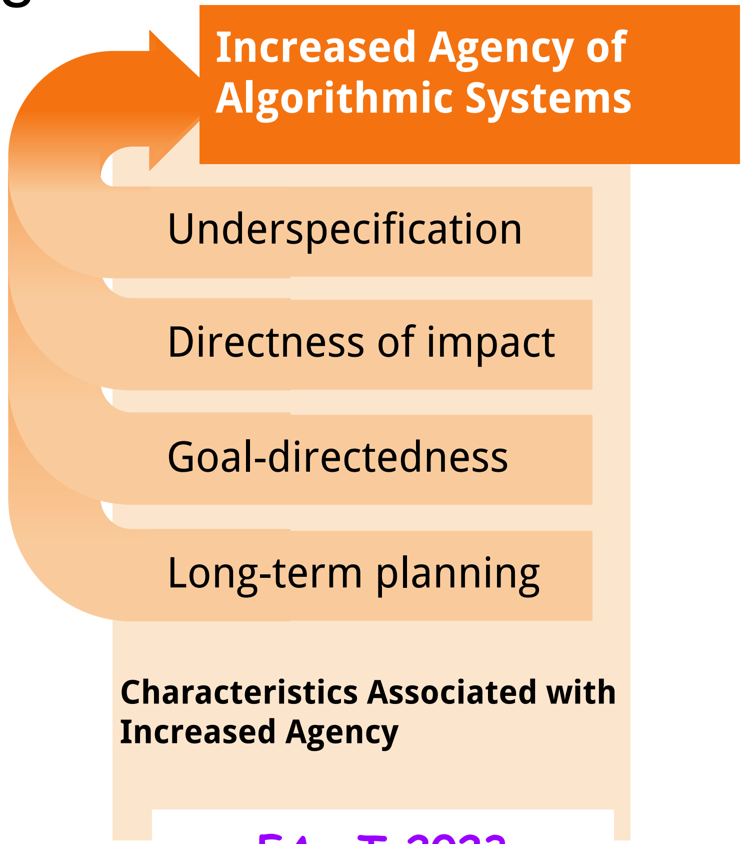
AI can't
do bad things

Motivation Control



AI doesn't want to
do bad things

# Agents




FAccT 2023

**Harms from Increasingly Agentic Algorithmic Systems**

ALAN CHAN[*,†,‡], Mila, Université de Montréal, Canada
REBECCA SALGANIK[†], Mila, Université de Montréal, Canada
ALVA MARKELIUS[†], University of Cambridge, UK
CHRIS PANG[†], University of Cambridge, UK
NITARSHAN RAJKUMAR[†], University of Cambridge, UK
DMITRII KRASHENINNIKOV[†], University of Cambridge, UK
LAURO LANGOSCO[†], University of Cambridge, UK
ZHONGHAO HE[†], University of Cambridge, UK
YAWEN DUAN[†], University of Cambridge, UK
MICAH CARROLL[†], University of California, Berkeley, USA
MICHELLE LIN, McGill University, Canada
ALEX MAYHEW, University of Western Ontario, Canada
KATHERINE COLLINS, University of Cambridge, UK
MARYAM MOLAMOHAMMADI, Mila, Canada
JOHN BURDEN, Center for the Study of Existential Risk, University of Cambridge, UK
WANRU ZHAO, University of Cambridge, UK
SHALALEH RISMANI, McGill University, Mila, Canada
KONSTANTINOS VOUDOURIS, University of Cambridge, UK
UMANG BHATT, University of Cambridge, UK
ADRIAN WELLER, University of Cambridge, UK
DAVID KRUEGER[*], University of Cambridge, UK
TEGAN MAHARAJ[*,†], University of Toronto, Canada

Existential safety trilemma:
**Choose 2** (maximum)

# Timelines and take-off speed

- Timelines: how soon?
- Take-off: how fast?
- Classic ideas:
  - “Recursive self-improvement”
  - FOOM
- Recent ideas:
  - Scaling
  - Automating AI R&D

# AI Policy

- “Evals”
- Compute governance
- Win the race, do a “pivotal act”

# Other

- Robustness
  - Edge cases, generalization, untestableness
  - Perverse instantiation, Goodharting
- Interpretability and assurance (safety cases, evals)

# Misconceptions and objections

# Misconceptions

- It’s all about timelines and take-off speeds (e.g. “AI2027”)
- It’s a big tech scam
- It’s a bunch of Luddites
- It’s about making LLMs “woke”
- …

# Objections

- It’s too soon to worry
- It’s too late to stop it
- AI is just another (normal) technology
- “I’m not worried about AI turning evil, I’m worried about people misusing AI”
- We’ll just build it safely
- Philosophical objections
  - How to deal with speculative risks?
  - How to forecast unprecedented events?
  - AI consciousness? “Carbon chauvinism”?
- …

**What are yours?**

# The AI Safety <u>Movement</u>

“TESCREAL”: Transhumanism, Extropianism, Singularitarianism, Cosmism, Rationalist ideology, Effective Altruism, and Longtermism

More nuanced:

## Building the Epistemic Community of AI Safety


Shazeda Ahmed
shazeda@g.ucla.edu
Center on Race and Digital Justice,
University of California - Los Angeles
California, USA

Klaudia Jazwinska
klaudia@princeton.edu
Center for Information Technology
Policy, Princeton University
New Jersey, USA

Archana Ahlawat
archana.ahlawat@princeton.edu
Center for Information Technology
Policy, Princeton University
New Jersey, USA

Amy Winecoff
aw0934@princeton.edu
Center for Information Technology
Policy, Princeton University
New Jersey, USA

Mona Wang
monaw@princeton.edu
Center for Information Technology
Policy, Princeton University
New Jersey, USA

# Actual properties of the movement:

- Lots of CS majors
- Lots of young people
- Lots of men
- Lots of Anglos
- Lots of left-libertarian politics
- Lots of transhumanists
- Lots of (aspiring) “Effective Altruists”
    - And “EA-adjacents”

# Mistakes were made... **which ones are we still making?**

- **Techno-solutionism**
- **Giving up on** governance / **global coordination**
- Acting like we're "Holding all the cards" / going it alone
- **Not engaging with academia enough**
- **Allying with big tech over civil society**
- Being too conciliatory to ML
  - Good cop / bad cop -- without the bad cop!
- **Meta-level:**
  - **Nepotistic funding**
  - Deferring to nebulous sense of community consensus
  - **Growing pains**

# Funders

- 3 big pots of money
  - Open Philanthropy (Facebook founder)
  - Survival and Flourishing Fund (Skype founder)
  - Future of Life Institute (Ethereum creator)
- Previously: FTX / Sam Bankman-Fried
- Also:
  - Schmidt Futures
  - Macroscopic Ventures
  - Longview Philanthropy
  - Various “high net worth individuals”

# Other related movements/communities

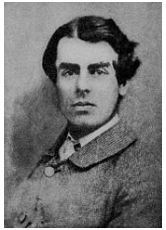
AI ethics

AI safety

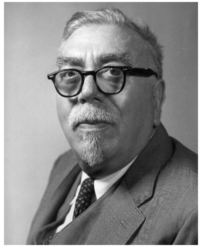
e/acc

# Pre/post 2023

# Timeline of AI x-risk



*"I fear none of the existing machines; what I fear is the extraordinary rapidity with which they are becoming something very different to what they are at present."* - ***Erewhon* (Samuel Butler, 1872)**



*"We had better be quite sure that the purpose put into the machine is the purpose which we really desire and not merely a colorful imitation of it"* - **Norbert Wiener, 1960**

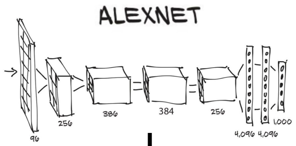
*"At some stage therefore we should have to expect the machines to take control"* - **Alan Turing, 1951**

*"There will be a strong and increasing pressure to improve AI up to human-level. If there is a way of guaranteeing that superior artificial intellects will never harm human beings then such intellects will be created. If there is no way to have such a guarantee then they will probably be created nevertheless."* - **Nick Bostrom, 1998**

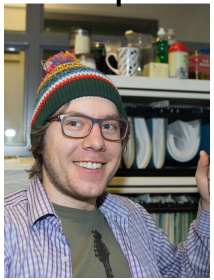
*"The AI does not hate you, nor does it love you, but you are made out of atoms which it can use for something else."* - **Eliezer Yudkowsky, 2006**

# Timeline of AI x-risk

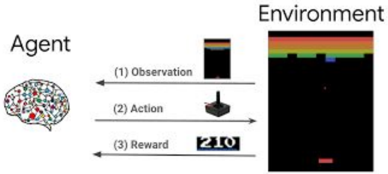
2012



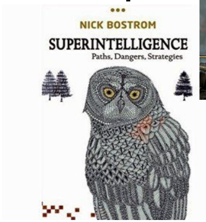

2013

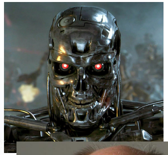


2013




2014



2016

Concrete Problems in AI Safety

Dario Amodei* Google Brain
Chris Olah* Google Brain
Jacob Steinhardt Stanford University
Paul Christiano UC Berkeley
John Schulman OpenAI
Dan Mané Google Brain

2016

"AI Alignment"

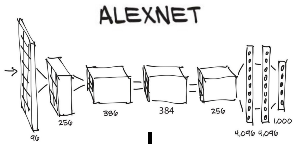
2020

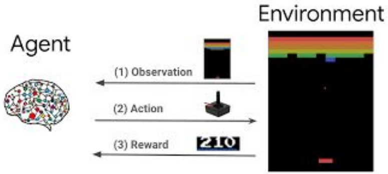
GPT3

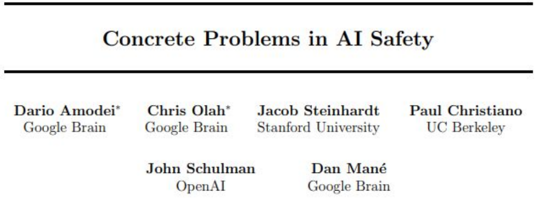
2021


future of life INSTITUTE

# Timeline of AI x-risk

Concrete Problems in AI Safety

Dario Amodei* Google Brain
Chris Olah* Google Brain
Jacob Steinhardt Stanford University
Paul Christiano UC Berkeley
John Schulman OpenAI
Dan Mané Google Brain

2012

2013

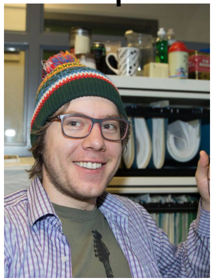
2016

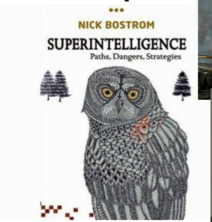
2021

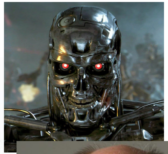
2013

2014

2020





My PhD

GPT3

Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war.

https://www.safe.ai/statement-on-ai-risk

**Geoffrey Hinton**
Emeritus Professor of Computer Science, University of Toronto

**Yoshua Bengio**
Professor of Computer Science, U. Montreal / Mila

**Demis Hassabis**
CEO, Google DeepMind

**Sam Altman**
CEO, OpenAI

**Dario Amodei**
CEO, Anthropic

**Dawn Song**
Professor of Computer Science, UC Berkeley

**Ted Lieu**
Congressman, US House of Representatives

**Bill Gates**
Gates Ventures

**Ya-Qin Zhang**
Professor and Dean, AIR, Tsinghua University

Center *for* AI Safety

# 2023

- ChatGPT → AI is mainstream
- Hinton, Bengio “safety-pilled” → CAIS statement → UK AISI
- Democratic administration → AI Ethics ascendent → “x-risk is a distraction”

# post-2023

- CEOs stop talking about x-risk
- Safety community schism
- Republican administration → e/acc ascendent → “Doomers”

# What I’m <u>Doing</u>

# Let’s wake the world up

***I want people to grasp, on a gut level, that AI could kill them and those they love — and to fill the world with clear signals of that reality.***

# Let’s wake the world up

***I want people to grasp, on a gut level, that AI could kill them and those they love — and to fill the world with clear signals of that reality.***

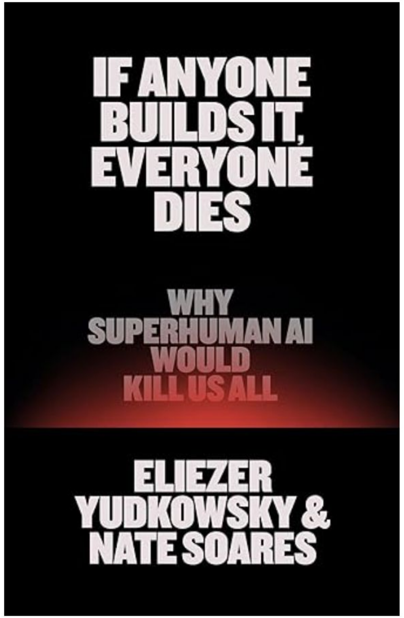
- Claim: at least 10% x-risk within 10 years.
- Obvious response: shut it down.
- Obvious question: but **how?**
- (non-)obvious answer: stop making advanced AI chips and “fabs”

# Goal: A Mass Movement Against AI Takeover

- Emphasize loss of control / societal-scale risk in a broad sense:
  - Gradual disempowerment, unemployment, concentration of power, the injustice of tech companies making this decision for all of humanity, etc.
- Non-partisan: this affects **everyone.**
- A gentle on-ramp: “something my parents would join”
- Current vision: a membership association
  - Educate and train members/organizers
  - Outreach within members existing communities
  - Building infrastructure for coordinate political action
  - Maintain strong message discipline
- **Lots of uncertainties still!**

# Existing work:

- NB: not a popular strategy! Cf [Don’t Build An AI Safety Movement - by Anton Leicht]
- Most work aims at **elites / policymakers**, not **public / civil society**
- Exceptions:



**Guido Reichstadter** @wolflovesmelon · Sep 4

Hi, my name's Guido Reichstadter, and I'm on hunger strike outside the offices of the AI company Anthropic right now because we are in an emergency. Anthropic and other AI companies are racing to create ever more powerful AI systems. These AI's are being used to inflict serious

Show more


645 530 1.7K 1M

# What <u>You</u> Can Do

*Personal opinion: Don’t inconvenience random people.*

**Guido Reichstadter** @wolflovesmelon · Sep 4

Hi, my name's Guido Reichstadter, and I'm on hunger strike outside the offices of the AI company Anthropic right now because we are in an emergency. Anthropic and other AI companies are racing to create ever more powerful AI systems. These AI's are being used to inflict serious

Show more


645 530 1.7K 1M

# A few ideas for how to contribute

- Conceptual work
  - What is power?
  - What is agency?
  - Understanding (dis)empowerment and what makes AI more/less tool-like
- Multi-agent systems and cooperation
  - Political economy of AI
  - Cf cooperative AI foundation
  - Algorithmic collusion
- Governance
  - Democratic/representative AI
  - What sort of positive visions are there for the future with(/out) AI?
  - Values, norms, and institutions for AI

# A few more ideas for how to contribute

- Forecasting and tracking trends:
    - Economic modelling
    - Economic and social indicators of AI risks, such as gradual disempowerment
        - Mass unemployment insurance? -- get a market signal!
- Which social systems will AI disrupt?
    - How/when/why?
    - Will this be good or bad?
    - Leveraging new capabilities for good
- Human/AI interaction
    - Mental health
    - (individual and large-scale) manipulation
    - Avoiding preference falsification

# A few more ideas for how to contribute

- Conceptual work
  - What is power?
  - Understanding (dis)empowerment and what makes AI more/less tool-like
- Multi-agent systems and cooperation
  - Political economy of AI
- Which social systems will AI disrupt?  How/when/why?  Will this be good or bad?
- Governance
  - Democratic/representative AI
    - What do people want and how can we tell?  What sort of positive visions are there for the future with(/out) AI?
  - Values, norms, and institutions for AI
- Human/AI interaction
  - Mental health
  - (individual and large-scale) manipulation
  - 
- Forecasting and tracking trends:
  - Economic modelling
  - Economic and social indicators of AI risks, such as gradual disempowerment